VARIANSI: Journal of Statistics and Its Application on Teaching and Research **ISSN 2684-7590 (Online)**
Vol. 7 No. 2 (2025), 133-142

DOI: 10.35580/variansiunm386

Klasifikasi Curah Hujan Di Kota Makassar Menggunakan *Gradient Boosting Machine* (GBM)

Hardianti Hafid, Zulkifli Rais*, Akhmad Rezky Ramadhana T

(Program Studi Statistika, Universitas Negeri Makassar, Indonesia)

Keywords: Curah Hujan, *Gradient Boosting Machine*, Klasifikasi, Machine Learning.

Abstract:

Rainfall is one of the important parameters in determining the climate of an area. Makassar, as one of the largest cities in Indonesia, has varying rainfall patterns throughout the year. This research aims to classify rainfall in Makassar City using the Gradient Boosting Machine (GBM) method. The secondary data used in this study were obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG), with predictor variables including wind speed, humidity, and air temperature, and the target variable being rainfall category, consisting of no rain, very light rain, light rain, moderate rain, heavy rain, and very heavy rain. To address class imbalance in the data, this study uses the Random Undersampling (RUS) technique. The GBM model with optimal hyperparameter configuration (n_estimators, learning_rate, max_depth, subsample, min_samples_leaf, max_features) achieved a classification accuracy rate of 98.46%, precision of 93%, recall of 98%, and F1-score of 95% with a training and testing data split of 80:20. The research results show that the GBM method is able to classify rainfall very well and can be used as a tool to assist in disaster mitigation planning and water resource management in Makassar City.

1. Pendahuluan

Data mining merupakan sebuah proses analisis yang dirancang untuk menemukan pola tersembunyi, korelasi dan tren yang bermanfaat (Larose & Larose, 2014). Data mining merupakan cabang dari bidang yang lebih besar dari *Knowledge Discovery in Databases (KDD)* yang bertujuan mengekstraksi pengetahuan dari data melalui Langkah-langkah interatif dan interaktif (Fayyad dkk, 1996).

Seiring dengan perkembangan teknologi, penerapan pembelajaran mesin *(machine learning)* menjadi semakin luas, tidak hanya untuk mengotomatisasi proses pengambilan Keputusan, tetapi juga untuk melakukan klasifikasi dan prediksi berbasis data (Domingos, 2012). Salah satu metode dalam *machine learning* adalah *decision tree*, yaitu model prediktif yang memvisualisasikan proses pengambilan Keputusan melalui struktur berbentuk pohon.

Selain *decision tree*, metode *ensemble learning* juga semakin banyak digunakan karena kemampuannya dalam menggabungkan kekuatan dari beberapa model untuk meningkatkan akurasi prediksi (Hochachka dkk, 2007). Salah satu teknik *ensemble* yang paling efektif dan banyak digunakan adalah *Gradient Boosting Machine (GBM)* yang terbukti unggul dalam berbagai kasus klasifikasi dan regresi (Chen & Guestrin, 2016).

* Corresponding author.
E-mail address: zulkifli.rais89@unm.ac.id

Berbagai studi menunjukkan bahwa teknik pembelajaran mesin *(machine learning)* dan metode *ensemble* telah digunakan secara luas, seperti pada pemetaan potensi mata air tanah (Mohammadi, 2021), prediksi radiasi matahari global (Mishra dkk, 2023), dan estimasi permintaan air di wilayah metropolitan (Shuang & Zhao, 2021). Hal ini menunjukkan bahwa penggunaan teknik pembelajaran mesin, terutama *GBM* dan *ensemble learning* telah memberikan kontribusi yang signifikan dalam pemetaan dan pemodelan fenomena alam seperti potensi air tanah, radiasi matahari global, dan aliran hujan.

Curah hujan merupakan salah satu parameter penting dalam menentukan iklim suau daerah. Kota Makassar sebagai salah satu kota terbesar di Indonesia memiliki pola curah hujan yang bervariasi sepanjang tahun. Oleh karena itu, penerapan teknik *data mining*, metode *ensemble*, dan *Gradient Boosting Machine* sangat relevan digunakan untuk mengoptimalkan analisis curah hujan di Kota Makassar untuk mendukung pengambilan keputusan yang lebih baik dalam manajemen sumber daya air dan mitigasi resiko terkait cuaca extrem.

2. Tinjauan Pustaka

2.1 Data Mining

Data mining adalah proses analitis yang digunakan untuk menemukan pola tersembunyi dalam data berukuran besar dengan memanfaatkan teknik dari berbagai disiplin ilmu, seperti kecerdasan buatan dan statistik (Ha dkk, 2011). Beberapa teknik umum dalam *data mining* yaitu klasifikasi, *clustering, association, regression, forecasting, sequence analysis,* dan *deviation analysis*.

2.2 Klasifikasi

Menurut (Pang-Ning Tan,Michael Steinbach,Vipin Kumar, 2019), klasifikasi merupakan proses kritis dalam analisis data yang digunakan dalam berbagai aplikasi, termasuk diagnosa medis, deteksi penipuan keuangan, dan klasifikasi teks,dengan kemampuannya untuk membuat prediksi berdasarkan pola-pola yang ada dalam data, klasifikasi memberikan kontribusi yang signifikan dalam pengambilan keputusan di berbagai domain.

2.3 Ensemble

Metode *ensemble* telah membuktikan keunggulannya dalam dunia machine learning dengan cara yang cukup signifikan. Salah satu keunggulan utama dari metode ensemble adalah kemampuannya untuk mengatasi masalah overfitting. Menggunakan pendekatan seperti Bagging atau Boosting, di mana beberapa model yang berbeda digabungkan, kita dapat mengurangi kesalahan yang disebabkan oleh variabilitas yang tinggi di dalam model individu. Ini sangat bermanfaat dalam situasi di mana data pelatihan terbatas atau rentan terhadap noise (Dietterich, 2000). Tidak hanya itu, metode ensemble juga dapat meningkatkan akurasi prediksi secara signifikan.

Salah satu algoritma ensemble learning yang populer adalah boosting. Boosting diperkenalkan oleh Robert E. Schapire pada tahun 1998 merupakan salah satu metode ensemble learning yang dapat meningkatkan kinerja dari beberapa hasil klasifikasi yang lemah agar dapat menjadi proses klasifikasi yang kuat. Teknik boosting dapat dilihat sebagai metode model rata-rata yang awalnya regresi (Syarif dkk, 2012).

2.4 Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) adalah sebuah metode *machine learning* yang termasuk dalam kategori *ensemble learning*. *GBM* bekerja dengan membangun serangkaian model prediksi secara berurutan, di mana setiap model baru berusaha memperbaiki kesalahan prediksi dari model sebelumnya. Proses ini dilakukan dengan mengoptimalkan fungsi kerugian tertentu, seperti *Mean Squared Error (MSE)* untuk masalah regresi atau *Cross Entropy* untuk masalah klasifikasi (Friedman, 2001).

Rumus dasar dari *GBM* dapat dijelaskan sebagai berikut:

a. Inisialisasi model-model awal $F_0(x)$ diinisialisasikan dengan nilai konstan,biasanya berupa nilai rata-rata dari target yang ingin di prediksi.

b. Iterasi(*boosting*) : Pada setiap iterasi model baru $h_m(x)$ ditambahkan untuk memperbaiki model sebelumnya. Model baru ini dipilih berdasarkan kemampuannya dalam mengurangi kesalahan residual dari model sebelumnya. Proses ini menghasilkan prediksi pada iterasi M seperti berikut (Friedman, 2001):

$$F_m(x) = F_{m-1}(x) + h_m(x) \quad \textbf{(2.1)}$$

dimana:

1. $F_m(x)$ = prediksi pada iterasi m
2. $F_{m-1}(x)$ = prediksi dari model pada iterasi sebelumnya
3. $h_m(x)$ = model baru yang ditambahkan pada iterasi m

c. Optimasi fungsi kerugian : *GBM* bekerja dengan mengoptimalkan fungsi kerugian tertentu,seperti *MSE* untuk Regresi dan *Cross Entropy* untuk Klasifikasi,agar meminimalkan kesalahan prediksi.

d. Penyelarasan : setelah jumlah iterasi yang ditentukan telah dicapai atau kriteria penghentian lainnya terpenuhi,prediksi akhir F(x) diperoleh dengan menambahkan semua model $h_m(x)$ yang telah di tambahkan selama iterasi (S. Raschka, 2019):

$$F_m(x) = F_0(x) + \sum_{m=1}^{M} h_m(x) \quad \textbf{(2.2)}$$

2.5 Akurasi

Akurasi adalah salah satu metrik evaluasi kinerja yang paling umum digunakan dalam data mining dan pembelajaran mesin, khususnya dalam konteks klasifikasi. Akurasi sangat penting karena memberikan gambaran keseluruhan tentang performa model dalam mengklasifikasikan data (S. Raschka, 2019). Rumus akurasi dinyatakan sebagai berikut (S. Raschka, 2019):

$$Akurasi = \frac{(TP+TN)}{(TP+TN+FP+FN)}\ x\ 100\% \quad \textbf{(2.3)}$$

Metrik evaluasi lain sering digunakan untuk melengkapi akurasi, seperti *Precision*, *Recall*, dan *F1-score*. *Precision* mengukur proporsi prediksi positif yang benar-benar positif, memberikan indikasi seberapa akurat prediksi positif yang dibuat oleh model (Ha dkk, 2011):

$$Precision = \frac{TP}{TP+FP}\ x\ 100\% \quad \textbf{(2.5)}$$

Recall, juga dikenal sebagai Sensitivity atau True Positive Rate, mengukur proporsi kasus positif yang benar-benar terdeteksi sebagai positif, menunjukkan seberapa baik model dalam menangkap semua kasus positif (Ha dkk, 2011):

$$Recall = \frac{TP}{TP+FN}\ x\ 100\ \% \quad \textbf{(2.6)}$$

F1-score adalah rata-rata harmonis dari Precision dan Recall, yang memberikan keseimbangan antara keduanya dan berguna dalam situasi di mana ada ketidakseimbangan kelas (Ha dkk, 2011):

$$F1 - Score = 2\ \times \frac{Precision \times Recall}{Precision+Recall} \quad \textbf{(2.7)}$$

Dalam rangka memahami bagaimana berbagai metrik evaluasi ini dihitung dan digunakan, penting untuk memperkenalkan konsep dari *Confusion Matrix*. *Confusion Matrix* adalah sebuah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan menunjukkan jumlah prediksi yang benar dan salah, baik untuk kelas positif maupun negatif. Tabel ini memberikan gambaran rinci tentang bagaimana model klasifikasi bekerja dalam memisahkan kelas-kelas yang berbeda dan membantu dalam menghitung metrik evaluasi seperti Akurasi, *Precision*, *Recall*, dan *F1-score* (Powers, 2020).

Tabel 2.1 *Confusion Matrix*

Nilai Aktual	***Negative Positive***	**Nilai Prediksi**	
		Positive	*Negative*
		True Positive (TP)	*False Negative (FN)*
		False Positive (FP)	*True Negative (TN)*

Sumber **:** (Powers, 2020)

2.6 Hyperparameter Tuning

Parameter pada metode *Gradient Bopsting Machine(GBM)* yang digunakan untuk *hyperparameter tuning* dapat dilihat pada Tabel 2.3 berikut ini (Friedman, 2001):

Tabel 1 Keterangan *Hyperparameter*

Hyperparameter	Keterangan
n_estimator	jumlah pohon keputusan (tree) yang akan dibuat dan digabungkan dalam model. Semakin banyak pohon, semakin kuat model, tapi bisa jadi lebih lambat.
learning_rate	seperti "kecepatan belajar" model. Semakin kecil nilainya, model belajar lebih pelan tapi hati-hati (mengurangi risiko overfitting). Biasanya di bawah
max_depth	Batas kedalaman setiap pohon. Semakin dalam pohon, semakin kompleks. Tapi pohon yang terlalu dalam bisa membuat model menghafal data (overfitting).
subsample	Persentase data yang diambil secara acak dari data pelatihan untuk membuat setiap pohon.
min_samples_split	Jumlah minimal data yang dibutuhkan agar suatu cabang pohon bisa dipecah lagi. Nilai besar mencegah pohon terlalu rumit.
min_samples_leaf	Jumlah minimal data yang harus ada di "ujung daun" pohon (bagian terakhir). Ini mencegah model membuat keputusan dari data yang terlalu sedikit.
max_features	Jumlah fitur (kolom dalam data) yang dipilih secara acak saat membuat pembagian di pohon. Ini membantu model lebih cepat dan lebih variatif.
random_state	Angka acak

2.7 Curah Hujan

Curah hujan adalah jumlah air dalam bentuk cair atau padat yang jatuh ke bumi dalam periode tertentu dan diukur dalam milimeter (mm). Curah hujan adalah salah satu indikator penting dalam studi iklim dan meteorologi, karena mempengaruhi berbagai aspek kehidupan manusia, termasuk pertanian, pengelolaan sumber daya air, dan mitigasi bencana alam seperti banjir dan kekeringan (Trenberth, 2011).

3. Metode Penelitian

3.1 Jenis Penelitian

Penelitian ini merupakan penelitian kuantitatif, dimana data curah hujan di Kota Makassar dikumpulkan kemudian diklasifikasi menggunakan metode *Gradient Boosting Machine (GBM)*.

3.2 Sumber Data

Data pada penelitian ini adalah data sekunder, yaitu data Curah Hujan yang Terjadi di Kota Makassar yang diambil dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) yang dapat diakses melalui: https://dataonline. bmkg. go. id/data_iklim

3.3 Definisi Operasional Variabel

1. Kecepatan Angin (X1)

Kecepatan angin mengacu pada kecepatan gerakan udara di sekitar suatu lokasi(Muhaniroh & Syech, 2021)

2. Kelembaban Udara (X2)

Kelembaban udara merujuk pada jumlah uap air yang terkandung dalam suatu volume udara. Hal ini mempengaruhi kenyamanan termal, pembentukan awan, dan kondisi cuaca secara umum (Septiani, 2024).

3. Suhu Udara (X3)

Suhu udara adalah ukuran dari tingkat panas atau dinginnya udara di suatu lokasi pada waktu tertentu (Septiani, 2024).

4. Curah Hujan (Y)

Curah hujan mengacu pada jumlah air yang jatuh ke permukaan bumi dalam bentuk hujan selama periode waktu tertentu (Muhaniroh & Syech, 2021).

3.4 Prosedur Penelitian

Adapun tahapan-tahapan yang dilakukan berdasarkan pada tujuan penelitian adalah sebagai berikut:
1. Melakukan pengambilan data pada situs website Badan Meteorologi, Klimatologi, dan Geofisika (BMKG).
2. Melakukan klasifikasi dengan menggunakan metode Gradient Boosting Machine (GBM)
3. Interpretasi hasil analisis.
4. Penarikan kesimpulan.
5. Menyusun laporan hasil penelitian.

3.5 Teknik Analisis Data

Adapun langkah-langkah Teknik analisis yang akan dilakukan dalam penelitian ini adalah:

1. Mengumpulkan data curah hujan dari website resmi Badan Meteorologi, Klimatologi, dan Geofisika (BMKG).
2. Melakukan pengecekan data hilang
3. Melakukan imputasi data
4. Memberikan Label Kategorikal
5. Membagi data latih dan data uji
6. Inisialisasi model GBM
7. Melatih model GBM menggunakan data latih.
8. Mengevaluasi model dengan menghitung akurasi klasifikasi dengan membandingkan hasil prediksi GBM dengan label sebenarnya pada data uji, lalu membentuk confusion matrix untuk mengevaluasi performa model secara lebih detail.

4. Hasil dan Pembahasan

4.1. Analisis Deskriptif

Berikut merupakan hasil analisis deskriptif untuk menjelaskan Gambaran umum variabel prediktor dan variabel respon:
1. Curah Hujan (Y)

Tabel 2. Frekuensi Curah Hujan di Kota Makassar Pada Periode 1 Januari 2020- 30 Juni 2024

Curah Hujan	Frekuensi	Persentase (%)
Tidak Hujan	704	43,24
Hujan Sangan Ringan	362	22,24
Hujan Ringan	297	18,24
Hujan Sedang	182	11,18
Hujan Lebat	60	3,69
Hujan Sangat Lebat	23	1,41
Total	**1628**	**100**

Dari Tabel di atas dapat diketahui bahwa dari total 1.620 hari amatan/data curah hujan,terdapat 704 (43,24%) hari yang berstatus tidak hujan,362 (22,24%) berstatus hujan sangat ringan,297 (18,24%) berstatus hujan ringan,182 (11,18%) berstatus hujan sedang,60 (3,69%) berstatus hujan lebat, dan 23 (1,41%) berstatus hujan sangat lebat.

Gambar 1. Frekuensi Curah Hujan di Kota Makassar Periode 1 Januari 2020-30 Juni 2024

Pada Gambar diatas menunjukan distribusi curah hujan pada sampel pengamatan. Curah hujan pada setiap klasifikasi memiliki distribusi data yang tidak seimbang.

2. Kecepatan Angin (X_1)

Tabel 3 Analisis Deskriptif Kecepatan Angin di Kota Makassar Pada Periode 1 Januari 2020-30 Juni 2024

Ukuran	Nilai (m/s)
Mean	4,75
Median	4
Modus	4
Minimum	2
Maksimum	17
Range	15
Standar deviasi	1.63
Variansi	2.66

Pada Tabel 4.2 disajikan beberapa ukuran yang menjelaskan deskripsi singkat dari variabel kecepatan angin di Kota Makassar antara 1 Januari 2020 hingga 30 Juni 2024. Rata-rata (mean) kecepatan angin di Kota Makassar adalah sebesar 4,75 m/s, dengan nilai median dan modus yang sama-sama sebesar 4 m/s. Kecepatan angin tercatat memiliki nilai minimum 2 m/s, yang menunjukkan bahwa pada beberapa waktu tertentu kecepatan angin bisa sangat rendah. Sebaliknya, kecepatan angin tercatat mencapai nilai maksimum sebesar 17 m/s, yang terjadi pada beberapa periode tertentu, dengan rentang (range) kecepatan angin sebesar 15 m/s.
Selanjutnya, standar deviasi dari kecepatan angin di Kota Makassar adalah sebesar 1,63 m/s, dan variansi sebesar 2,66 m^2/s^2. Nilai ini menunjukkan bahwa data pengamatan kecepatan angin di Kota Makassar tersebar cukup luas dari rata-ratanya, dengan beberapa periode mengalami fluktuasi kecepatan angin yang signifikan.

3. Kelembapan Udara (X_2)

Tabel 4 Analisis Deskriptif Kelembapan Udara di Kota Makassar Pada Periode 1 Januari 2020-30 Juni 2024

Ukuran	Nilai (%)
Mean	79.7
Median	80
Modus	78
Minimum	54
Maksimum	96
Range	42
Standar deviasi	6,64
Variansi	44,07

Pada Tabel 4.3 disajikan beberapa ukuran yang menjelaskan deskripsi singkat dari variabel kelembapan udara di Kota Makassar antara 1 Januari 2020 hingga 30 Juni 2024. Rata-rata (mean) kelembapan udara di Kota Makassar adalah sebesar 79,7%, dengan nilai median yang hampir sama yaitu 80%. Modus dari kelembapan udara tercatat pada nilai 78%, yang menunjukkan bahwa angka ini paling sering muncul dalam data. Kelembapan udara memiliki nilai minimum sebesar 54%, yang menunjukkan kondisi kelembapan yang sangat rendah pada beberapa waktu tertentu, dan nilai maksimum sebesar 96%, menunjukkan kelembapan udara yang sangat tinggi pada periode lainnya, dengan rentang (range) kelembapan udara sebesar 42%.
Selanjutnya, standar deviasi kelembapan udara adalah sebesar 6,64%, dan variansi sebesar 44,07%. Nilai ini menunjukkan bahwa data kelembapan udara di Kota Makassar cenderung tersebar dengan cukup lebar di sekitar nilai rata-rata, dengan variasi kelembapan yang cukup signifikan antar waktu.

4. Suhu Udara (X_3)

Tabel 5 Analisis Deskriptif Suhu Udara
di Kota Makassar Pada Periode 1 Januari 2020-30 Juni 2024

Ukuran	**Nilai (**°C)
Mean	28,06
Median	28,2
Modus	28,7
Minimum	23,4
Maksimum	32,3
Range	8,9
Standar deviasi	1,15
Variansi	1,32

Pada Tabel 4.4 disajikan beberapa ukuran yang menjelaskan deskripsi singkat dari variabel suhu udara di Kota Makassar antara 1 Januari 2020 hingga 30 Juni 2024. Rata-rata (mean) suhu udara tercatat sebesar 28,06°C, dengan nilai median yang sedikit lebih tinggi yaitu 28,2°C. Modus suhu udara berada pada nilai 28,7°C, yang menunjukkan suhu ini paling sering muncul dalam data. Suhu udara di Kota Makassar memiliki nilai minimum sebesar 23,4°C, yang menunjukkan kondisi suhu yang relatif rendah pada beberapa waktu tertentu, dan nilai maksimum sebesar 32,3°C, yang menunjukkan suhu udara yang tinggi pada periode lainnya. Rentang (range) suhu udara ini sebesar 8,9°C.
Selanjutnya, standar deviasi suhu udara adalah sebesar 1,15°C, dan variansi sebesar 1,32°C. Nilai standar deviasi ini menunjukkan bahwa suhu udara di Kota Makassar cenderung tidak terlalu tersebar jauh dari rata-rata, dengan variasi sebesar 1,32°C. Nilai standar deviasi ini menunjukkan bahwa suhu udara di Kota Makassar cenderung tidak terlalu tersebar jauh dari rata-rata, dengan variasi suhu yang lebih terpusat.

4.2. Pembagian Data Latih dan Data Uji

Pada penelitian ini, dua skenario pembagian data digunakan: pertama, pembagian dengan proporsi 70:30 di mana 70% data digunakan untuk training dan 30% untuk testing, dan kedua, pembagian dengan proporsi 80:20, di mana 80% data digunakan untuk training dan 20% digunakan untuk testing. Pembagian ini dilakukan untuk menguji seberapa baik model GBM dalam menggeneralisasi data dan meningkatkan akurasi prediksi.

4.3. Undersampling Data Imbalance

Pada Tabel 4.1 dapat dilihat bahwa terdapat ketimpangan yang tinggi antara jumlah data pada status curah hujan normal dengan kasus hujan tinggi sehingga menyebabkan distribusi data tidak seimbang (*imbalance*). Dalam analisis klasifikasi, kasus data imbalance adalah hal yang umum dijumpai (Gumelar dkk, 2021).
Dalam mengatasi kasus data yang tidak seimbang, terdapat beberapa teknik yang dapat digunakan diantaranya yang sederhana adalah teknik *random undersampling (RUS)* dan *random oversampling (ROS)*. Dalam penelitian ini akan digunakan teknik RUS untuk mengatasi kasus data *imbalance* karena menghasilkan performa yang lebih baik dari teknik ROS. Berikut hasil penggunaan teknik RUS:

Tabel 6 Hasil Penggunaan RUS pada Data

Proporsi	**Frekuensi**	**Persentase (%)**

80:20	108	16,67
70:30	96	16,67

4.4. Klasifikasi Gradient Boosting Machine (GBM)

a. Analisis Klasifikasi Gradient Boosting Machine (GBM)

Langkah pertama yang dilakukan adalah melakukan hyperparameter tuning menggunakan fungsi pada *scikit-learn* yaitu *gridsearchcv* Untuk Mencari kombinasi nilai *hyperparameter* terbaik sebagai model akhir. Kombinasi nilai *hyperparameter Gradient Boosting* Machine (GBM) dapat dilihat pada Tabel 4.6 berikut ini:

Tabel 7 Kombinasi Nilai *Hyperparameter*

Hyperparameter	***Grid Search Values***
n_estimator	[50, 100, 200]
learning_rate	[0.01, 0.05, 0.1]
max_depth	[5, 10, 15]
subsample	[0.8, 1.0]
min_samples_split	[20, 50, 100]
min_samples_leaf	[10, 20, 30]
max_features	['sqrt', 'log2']

Dalam proses tuning parameter, metode seperti GridSearchCV dan *RandomizedSearchCV* sering digunakan untuk menemukan kombinasi parameter yang memberikan akurasi terbaik. Selain itu, evaluasi dilakukan dengan teknik validasi silang (cross-validation) untuk memastikan bahwa hasil yang diperoleh tidak hanya berlaku pada satu subset data tertentu, tetapi juga dapat diterapkan secara umum pada data lain (Géron dkk., 2019).
Penelitian ini menggunakan cv = 5 yang digunakan untuk mengevaluasi kinerja model sebanyak lima kali perulangan dalam proses *grid search* dari setiap parameter untuk mendapatkan nilai parameter terbaik. Parameter terbaik dari hasil *hyperparameter tuning Gradient Boosting Machine* (GBM) dapat dilihat pada tabel berikut:

Tabel 8 Nilai *Hyperparameter* Terbaik

Hyperparameter	**Grid Search Values**
n_estimator	200
learning_rate`	0.05
max_depth	5
subsample	1.0
min_samples_split	50
min_samples_leaf	10
max_features	Sqrt

Pada Tabel 4.7, dapat diketahui bahwa jumlah pohon yang digunakan dalam model adalah 200. Nilai *learning rate* yang digunakan untuk mengontrol penyusutan ukuran dalam proses pembelajaran ditetapkan sebesar 0.05. Kedalaman maksimum setiap pohon (*max_depth*) yang digunakan adalah 5, yang bertujuan untuk menghindari *overfitting*. Fraksi sampel yang digunakan dalam setiap iterasi *boosting* (*subsample*) adalah 1.0, yang berarti semua data pelatihan digunakan dalam setiap iterasi. Jumlah minimum sampel yang dibutuhkan untuk membagi suatu node (*min_samples_split*) adalah 50, sementara jumlah minimum sampel dalam satu daun pohon (*min_samples_leaf*) adalah 10. Selain itu, jumlah maksimum fitur yang digunakan untuk membangun setiap pohon (*max_features*) adalah akar dari jumlah total fitur (*sqrt*), yang membantu meningkatkan efisiensi dan mengurangi kemungkinan overfitting dengan tingkat akurasi grid search cv sebesar 0.962771.

b. Evaluasi Model

Tabel 9. Hasil Pengujian Pada Setiap Proporsi Data

Proporsi	**Akurasi**	***Precision***	***Recall***	***F1-Score***
70:30	97,95%	94%	98%	96%

80:20	**98,77%**	**95%**	**98%**	**96%**

Berdasarkan hasil perhitungan metrik evaluasi model, akurasi yang diperoleh pada proporsi data 70:30 adalah 97,95%, sedangkan pada proporsi 80:20 mencapai 98,77%. Dari perbandingan ini, dapat disimpulkan bahwa model dengan proporsi 80:20 memiliki akurasi yang lebih tinggi dibandingkan dengan 70:30.

Berdasarkan hasil klasifikasi yang dilakukan, data telah dikelompokkan berdasarkan variabel kecepatan angin, kelembapan udara, suhu udara, dan curah hujan. Hasil klasifikasi tersebut disajikan pada Tabel 4.10, yang menunjukkan bagaimana model Gradient Boosting Machine (GBM) mengelompokkan data ke dalam kategori tertentu berdasarkan karakteristik cuaca yang diamati.

Tabel 10 Hasil Klasifikasi Menggunakan *Gradient Boosting Machine* (GBM)

Kecepatan Angin	Kelembapan Udara	Suhu Udara	Curah Hujan	Hasil Klasifikasi
6	82	26,6	36	3
6	86	26,2	2,4	1
5	82	27,4	0,8	1
7	87	27,6	28	3
7	90	25,8	20,6	3
5	82	27,6	17,8	2
4	77	28,3	0	0
4	83	27,4	0	0
6	86	26,7	34,2	3
5	84	27	10,4	2

Berdasarkan hasil klasifikasi yang dilakukan menggunakan Gradient Boosting Machine (GBM), model ini mampu mengelompokkan data berdasarkan variabel kecepatan angin, kelembapan udara, suhu udara, dan curah hujan dengan cukup baik.

Dari hasil klasifikasi, terlihat bahwa curah hujan memiliki pengaruh yang cukup besar terhadap hasil klasifikasi. Semakin tinggi curah hujan, semakin besar kemungkinan data diklasifikasikan ke dalam kelas yang lebih tinggi. Hal ini dapat dilihat dari pola klasifikasi, di mana curah hujan yang lebih tinggi umumnya masuk ke dalam kelas 3, sedangkan curah hujan rendah atau nol cenderung masuk ke kelas 0 atau 1.

Selain itu, kelembapan udara juga menunjukkan keterkaitan dengan hasil klasifikasi, terutama pada kasus dengan curah hujan yang tinggi. Semakin tinggi kelembapan udara, semakin besar kemungkinan data masuk ke kelas yang lebih tinggi. Namun, kecepatan angin dan suhu udara tampaknya tidak memiliki pengaruh yang dominan dalam menentukan kelas.

Secara keseluruhan, hasil ini menunjukkan bahwa model GBM dapat menangkap pola hubungan antara variabel cuaca dengan hasil klasifikasi yang diberikan. Dengan demikian, model ini dapat digunakan untuk memprediksi kategori curah hujan berdasarkan parameter lingkungan lainnya dengan cukup akurat.

5. Kesimpulan

1. Untuk hasil evaluasi metode Gradient Boosting Machine (GBM) menghasilkan performa yang baik model menghasilkan nilai akurasi sebesar 98,46%, presisi sebesar 95%, recall sebesar 98%, F1 – Score sebesar 96% sehingga dapat diartikan bahwa kasus kalsifikasi curah hujan di kota Makassar dapat di klasifikasikan dengan baik atau tepat menggunakan metode Gradient Boosting Machine.
2. Untuk klasifikasi menggunakan metode Gradient Boosting Machine (GBM) dengan melakukan hyperparameter tunning menggunakan gridsearchCV di dapatkan nilai kombinasi Parameter terbaik yaitu n_estimators=200, learning_rate=0.1, max_depth=5, subsample=1.0, min_samples_split=50, min_samples_leaf=10, max_features="sqrt" dan parameter lain di atur secara default.

References

BMKG. (2008). *Curah Hujan dan Potensi Bencana Gerakan Tanah*. 1–7.

Badan Meteorologi, Klimatologi, dan G. (BMKG). (2023). *Buletin Hujan Bulanan – Updated Desember 2023 - Buletin IklimNo Title*. https://www.bmkg.go.id/iklim/buletin-iklim/buletin-hujan-bulanan-updated-desember-2023

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu, 785–794. https://doi.org/10.1145/2939672.2939785

Dietterich, T. (2000). *Ensemble Methods in Machine Learning BT - Lecture Notes in Computer Science. Lecture Notes in Computer Science*, *1857*(Chapter 1), 1–15. http://dx.doi.org/10.1007/3-540-45014-9_1%5Cnpapers2://publication/doi/10.1007/3-540-45014-9_1

Domingos, P. (2012*). A few useful things to know about machine learning. Communications of the ACM*, 55(10), 78–87. https://doi.org/10.1145/2347736.2347755

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From data mining to knowledge discovery in databases. AI Magazine*, 17(3), 37–53. https://doi.org/10.1609/aimag.v17i3.1230

Friedman, J. H. (2001*). Greedy function approximation: A gradient boosting machine. Annals of Statistics*, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Ha, J., Kambe, M., & Pe, J. (2011). *Data Mining: Concepts and Techniques. In Data Mining: Concepts and Techniques.* https://doi.org/10.1016/C2009-0-61819-5

Hochachka, wesley m., Caruana, R., Fink, D., Munson, A., Riedewald, M., Sorokina, D., & Kelling, S. (2007). *Data-Mining Discovery of Pattern and Process in Ecological Systems. The Journal of Wildlife Management*, 71(7), 2427–2437. https://doi.org/10.2193/2006-503

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data. Discovering Knowledge in Data*. https://doi.org/10.1002/9781118874059

Mishra, D. P., Jena, S., Senapati, R., Panigrahi, A., & Salkuti, S. R. (2023*). Global solar radiation forecast using an ensemble learning approach. International Journal of Power Electronics and Drive Systems*, 14(1), 496–505. https://doi.org/10.11591/ijpeds.v14.i1.pp496-505

Mohammadi, B. (2021). *A review on the applications of machine learning for runoff modeling. Sustainable Water Resources Management*, 7(6), 1–11. https://doi.org/10.1007/s40899-021-00584-y

Muhaniroh, M., & Syech, R. (2021). Analisis Pengaruh Suhu Udara, Curah Hujan, Kelembaban Udara Dan Kecepatan Angin Terhadap Arah Penyebaran Dan Akumulasi Particulate Matter (Pm10): Studi Kasus Kota Pekanbaru. *Komunikasi Fisika Indonesia*, *18*(1), 48. https://doi.org/10.31258/jkfi.18.1.48-57

Pang-Ning Tan,Michael Steinbach,Vipin Kumar, A. K. (2019). *Introduction to Data Mining eBook: Global Edition*. In Pearson Education Limited. https://doi.org/10.1016/b978-155558242-5/50003-6

Powers, D. M. W. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. 37–63. http://arxiv.org/abs/2010.16061

Shuang, Q., & Zhao, R. T. (2021). *Water demand prediction using machine learning methods: A case study of the beijing–tianjin–hebei region in china. Water (Switzerland)*, 13(3), 1–16. https://doi.org/10.3390/w13030310

Syarif, I., Zaluska, E., Prugel-Bennett, A., & Wills, G. (2012). Application of bagging, boosting and stacking to intrusion detection. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7376 LNAI, 593–602. https://doi.org/10.1007/978-3-642-31537-4_46

S. Raschka, V. M. (2019). Python Machine Learning (3rd ed.).

Septiani, N. (2024). *Pengaruh Suhu, Kelembaban Udara Terhadap Prediksi Curah Hujan Dan Relevansi Pada Fenomena Hujan Es Di Bandar Lampung*.

Trenberth, K. E. (2011). Changes in precipitation with climate change. Climate Research, 47(1–2), 123–138. https://doi.org/10.3354/cr00953